The MUC7$_\mathcal{T}$ Corpus

Katrin Tomanek and Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
{katrin.tomanek|udo.hahn}@uni-jena.de

1 Introduction

This document gives a short description of the creation of the MUC7$_\mathcal{T}$ corpus, its package structure, and the underlying data format. Finally, two use cases of MUC7$_\mathcal{T}$ are briefly described.

2 Creation of Muc7$_\mathcal{T}$

MUC7$_\mathcal{T}$ is an extension of the MUC7 corpus (Linguistic Data Consortium, 2001), where we couple common named entity annotation metadata with a time stamp which indicates the time measured for the linguistic decision making process.[1] Therefore, we ran a re-annotation initiative which targeted the named entity annotations (ENAMEX) of the English part of the MUC7 corpus, *viz.* PERSONS, LOCATIONS, and ORGANIZATIONS. The annotation was done by two advanced students of linguistics with good English language skills. For consistency reasons, the original guidelines of the MUC7 named entity task were used.

2.1 Data

The original MUC7 corpus consists of three distinct document sets for the named entity task (TRAIN, TEST, and DRY RUN). We used the TEST set to train the annotators and to develop the annotation design. The MUC7$_\mathcal{T}$ corpus consists of the articles from the TRAIN set which comprises 100 articles reporting on airplane crashes. We had to split lengthy documents (27 out of the 100) into halves so that they fitted in the screen of the annotation GUI without the need for scrolling.[2] Still, we had to exclude the following

[1] These time stamps should not be confounded with the annotation of temporal expressions (e.g., TIMEX in MUC7).

[2] We aimed at avoiding scrolling in order to keep the "mechanical" overhead of the actual annotation procedure as low as possible so that the annotation times would reflect basically the cognitive processes, only.

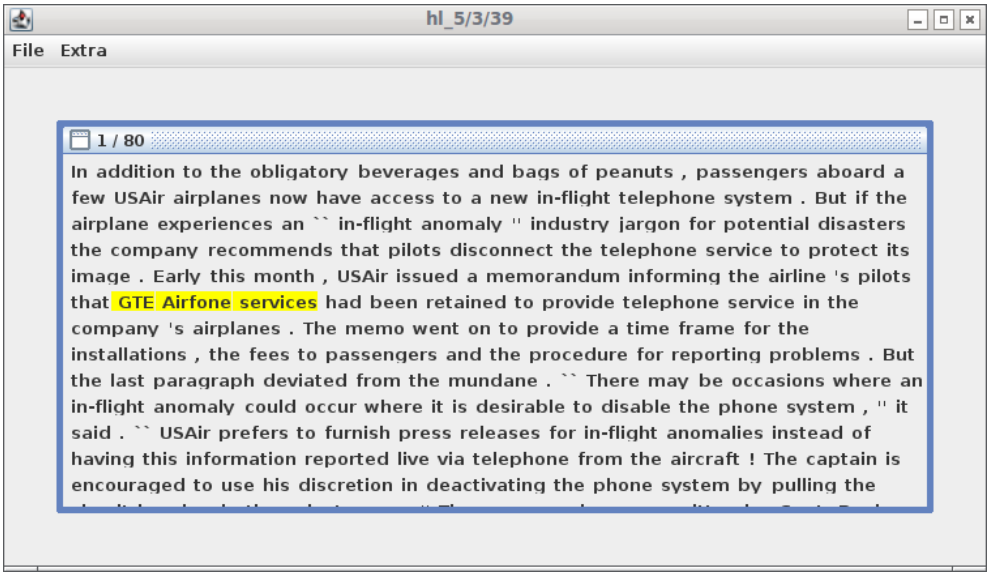

Figure 1: Screenshot of the annotation GUI showing an annotation example where the complex noun phrase "*GTE Airfone services*" is highlighted for annotation.

two documents due to extreme over-length which would have required overly many splits: `nyt960718.0792` and `nyt960721.0140`.

2.2 Annotation Principles

In the MUC7$_\mathcal{T}$ corpus, annotation time measurements are recorded for two syntactically different **annotation units**: *(a)* complete sentences and *(b)* complex noun phrases (CNPs) which are top-level noun phrases in the constituency structure of the respective sentence. The annotation task was defined such as to assign an entity type label to each token of an annotation unit. Please refer to Tomanek and Hahn (2010) for a discussion why CNPs were used and how these were derived automatically.

For precise time measurements, single **annotation examples** were shown to the annotators, one at a time. An annotation example consists of the chosen MUC7 document with one annotation unit (sentence or CNP) selected and highlighted (yet, without annotation). Only the highlighted part of the document could be annotated and the annotators were asked to read only as much of the visible context surrounding the annotation unit as necessary to make a proper annotation decision. Figure 1 shows a screenshot of the annotation GUI.

Annotation was performed in **blocks** of 500 CNP-level or 100 sentence-level annotation examples. In the MUC7$_\mathcal{T}$ corpus, annotation time meta data of both annotators is available on the CNP- and the sentence-level.

Further details on the creation and annotations principles of MUC7$_\mathcal{T}$ can be found in Tomanek and Hahn (2010). In this paper, you will also find an analysis of the inter-annotator agreement (Cohen's Kappa is $\kappa \approx 0.95$ for both annotators) as well as other corpus statistics.

3 Package Structure

Figure 2 shows the directory structure of the MUC7$_\mathcal{T}$ package. Some very short descriptions and remarks on each subdirectory:

- `data`

  This directory contains the actual MUC7$_\mathcal{T}$ data. You will find the data for annotator A and B, each separately. For both annotators, there is a version of MUC7$_\mathcal{T}$ with CNP-level and with sentence-level annotations. Section 4 discusses the used XML format in more detail.

- `docs`

  Contains this documentation as well as publications describing applications of MUC7$_\mathcal{T}$. There is also a small JavaDoc for the Java tools (see below).

- `dtd`

  You will find the Document Type Definition (DTD) for the data here.

- `tools`

  There is a small Java API which allows to read the MUC7$_\mathcal{T}$ XML data so that each annotation example is represented by a Java object. Besides the source code, you will also find a `jar` package. The code has been tested with Java 1.5 and Java 1.6.

4 Data Format

The MUC7$_\mathcal{T}$ corpus is stored in XML format. See Figure 3 for the respective DTD. There is an element `anno_example` for each annotation example. It has the original MUC7 document as text context. The MUC7 document was tokenized using the Stanford Tokenizer[3] with white spaces marking token boundaries. The following attributes are used for the element `anno_example`:

- `anno_time`: The time it took to annotate the annotation unit of this annotation example (time in milliseconds).

[3]The tokenizer is part of the Stanford Parser package which can be obtained from `http://nlp.stanford.edu/software/lex-parser.shtml`.

```
|-- data
|   |-- annotatorA
|   |   |-- CNPs
|   |   `-- sents
|   `-- annotatorB
|       |-- CNPs
|       `-- sents
|-- docs
|   `-- publications
|   `-- JavaDoc
|-- dtd
`-- tools
    `-- src
        `-- de
            `-- julielab
                `-- muc7timed
```

Figure 2: Directory structure of the MUC7$_\mathcal{T}$ package.

- `anno_unit_tokens`: All tokens of the annotation unit.
- `anno_unit_offset`: Offsets for the tokens of the annotation unit relative to all tokens in the annotation example.
- `anno_unit_labels`: Labels for the tokens of the annotation unit (these labels are taken from MUC7).
- `doc_id`: ID of the document of the annotation example.
- `sent_id`: ID of the sentence of the annotation example.
- `anno_unit_id`: ID of the unit of the annotation example. All three ids (`doc_id`, `sent_id`, and `anno_unit_id`) jointly yield a unique identifier of this annotation example. Moreover, they allow to regroup or reorder the annotation examples, e.g., by document or sentence. They can also be used as links between the CNP-level and the document-level version of MUC7$_\mathcal{T}$.
- `muc7_org_filename`: The name of the original MUC7 document from which this annotation example is taken.

Figure 4 shows such an annotation example in XML format. It stems from the original MUC7 file `nyt960721.0261`. It took about 4.5 seconds to annotate. The annotation unit (a CNP, here) consists of the 4th to 11th token (we start counting from 0) of the annotation example text; these tokens are *"the crash of Pan Am 103 over Lockerbie"* and *"Pan Am"* is marked as an organization and *"Lockerbie"* as a location. This annotation unit

```
<!ELEMENT anno_examples (anno_example)+>
<!ELEMENT anno_example (#PCDATA)>
<!ATTLIST anno_example
  anno_time CDATA #REQUIRED
  anno_unit_tokens CDATA #REQUIRED
  anno_unit_offset CDATA #REQUIRED
  anno_unit_labels CDATA #REQUIRED
  doc_id CDATA #REQUIRED
  sent_id CDATA #REQUIRED
  anno_unit_id CDATA #REQUIRED
  muc7_org_filename CDATA #REQUIRED
>
```

Figure 3: Document Type Definition (DTD) of the XML format.

was taken from the 68th MUC7 document, is in the 1st sentence of this document, and therein is the 1st CNP. The next CNP-level annotation example of this example would be for the annotation unit *"Scotland"* having `anno_unit_id="2"` (`doc_id` and `sent_id` would stay the same).

5 Use Cases of Muc7$_\mathcal{T}$

We created MUC7$_\mathcal{T}$ focusing on two main purposes both in the context of resource- and cost-conscious annotation strategies. On the one hand, it can be used for evaluations of selective sampling strategies, such as Active Learning (Cohn et al., 1996) – instead of empirically questionable assumptions on the necessary annotation efforts (e.g., the assumption of the uniformity of annotation costs over the number of linguistic units, typically tokens, to be annotated), MUC7$_\mathcal{T}$ now allows to run repeatable simulations on selective sampling strategies where the annotation effort can be expressed by the actual time needed to annotate a selected item. This use case is described in more detail in Tomanek and Hahn (2010) and Tomanek (2010).

Another use case for MUC7$_\mathcal{T}$ is the creation of predictive models for annotation costs. Such models are needed when selective sampling strategies, such as Active Learning, should not only select on the basis of estimated informativeness or utility of an example (to be maximized), but also taking into account the estimated time this example would require for annotation (to be minimized). As annotation costs are not known prior to annotation, their quantity has to be estimated. In Tomanek et al. (2010), we describe an empirical study where the annotators' reading behavior was observed with an eye-tracking device while a corpus was annotated. With the insights on factors influencing annotation time which we gathered through this study, we were able to induce such a much needed predictive model of annotation costs.

```
<anno_example
 anno_time="4448"
 anno_unit_labels="O O O ORGANIZATION ORGANIZATION O O LOCATION"
 anno_unit_offset="4,5,6,7,8,9,10,11"
 anno_unit_tokens="the crash of Pan Am 103 over Lockerbie"
 doc_id="68"
 sent_id="2"
 anno_unit_id="1"
 muc7_org_filename="nyt960721.0261" >

 MORICHES , N.Y. After the crash of Pan Am 103 over Lockerbie ,
 Scotland , in December 1988 , it took investigators seven days
 to determine that the cause was a bomb . But after a Boeing 737
 crashed on approach to Pittsburgh in September 1994 which was

 [...]
</anno_example>
```

Figure 4: A sample annotation example in XML format.

Acknowledgements

The construction of the MUC7$_\mathcal{T}$ corpus was partially funded by the BOOT-Strep Project (FP6-028099) within the 6th Framework Programme of the European Commission.

References

[Cohn et al. 1996] COHN, David; GHAHRAMANI, Zoubin; JORDAN, Michael: Active Learning with Statistical Models. In: *Journal of Artificial Intelligence Research* 4 (1996), pp. 129–145.

[Linguistic Data Consortium 2001] LINGUISTIC DATA CONSORTIUM: *Message Understanding Conference (MUC) 7*. LDC2001T02. FTP FILE. 2001. – Philadelphia: Linguistic Data Consortium.

[Tomanek 2010] TOMANEK, Katrin: *Resource-Aware Annotation through Active Learning*, Technical University of Dortmund, Dissertation, 2010.

[Tomanek and Hahn 2010] TOMANEK, Katrin; HAHN, Udo: Annotation Time Stamps Temporal Metadata from the Linguistic Annotation Process. In: *LREC'10: Proceedings of the 7th International Conference on Language Resources and Evaluation*, European Language Resources Association, 2010.

[Tomanek et al. 2010] Tomanek, Katrin; Hahn, Udo; Lohmann, Steffen; Ziegler, Jürgen: A Cognitive Cost Model of Annotations Based on Eye-Tracking Data. In: *ACL'10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010.